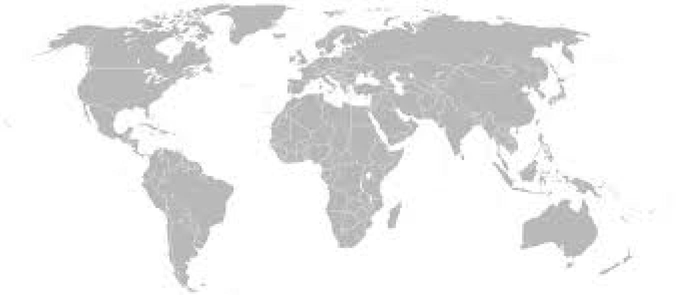
ITS 2.0 Automated Translation of Natural Language Content

MultilingualWeb Linguistic Technology

ITS2

London, 29 November 2013

Pedro Luis Díez Orzas



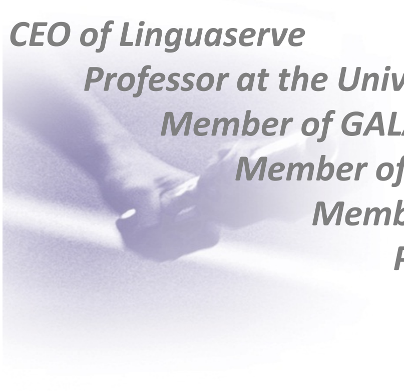
Translating and the Computer Conference

aslib
MANAGING INFORMATION
Est. 1924

linguaserve

Pedro Luis Díez Orzas

pedro.diez@linguaserve.com



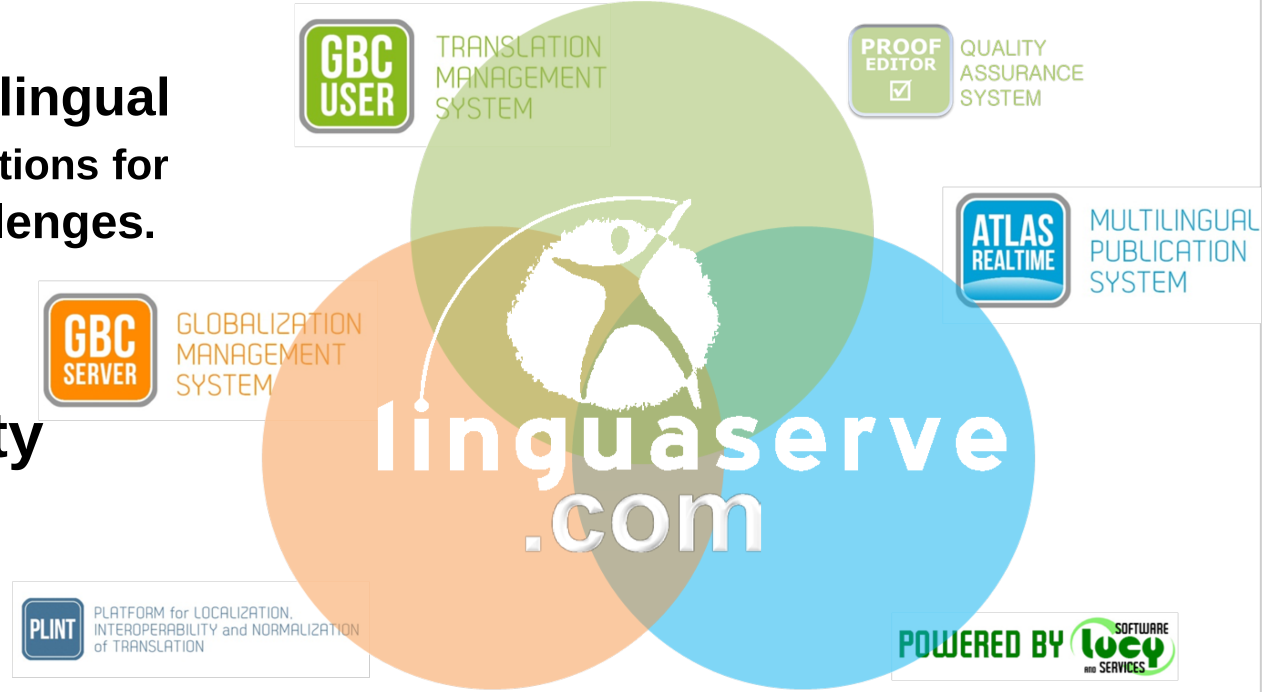
CEO of Linguaserve
Professor at the Universidad Complutense de Madrid
Member of GALA
Member of W3C MLW-LT and ITS IG
Member of the TNC-191 AENOR (Terminology)
PhD in Computational Linguistics

Translation, linguistic services
and cutting edge solutions

Linguaserve

specializes in multilingual web advanced solutions for 21st Century Challenges.

Experience in **interoperability** since 2002 and **real-time multilingual web publishing** since 2008.


GBC USER
TRANSLATION MANAGEMENT SYSTEM
PROOF EDITOR ☑
QUALITY ASSURANCE SYSTEM
ATLAS REALTIME
MULTILINGUAL PUBLICATION SYSTEM
GBC SERVER
GLOBALIZATION MANAGEMENT SYSTEM
linguaserve .com
PLINT
PLATFORM for LOCALIZATION, INTEROPERABILITY and NORMALIZATION of TRANSLATION
POWERED BY lucy SOFTWARE AND SERVICES

CONTENTS

01 Introduction

02 Standards, they are great…

03 Multilingual Web LT and ITS 2.0

04 ITS 2.0 interoperability CMS/TMS

05 Online MT Internationalisation

06 Conclusions

CONTENTS

01 Introduction

02 Standards, they are great…

03 Multilingual Web LT and ITS 2.0

04 ITS 2.0 interoperability CMS/TMS

05 Online MT Internationalisation

06 Conclusions

Why ITS 2.0?

- ITS 2.0 is a conceptual system of elements and attributes for the internationalization, translation and localization of web content.
- ITS 2.0 is not merely a tagging or labelling standard.
- ITS 2.0 can be represented in different formats.
- ITS 2.0 success is expected to materialize in real-life implementations (currently 20).
- ITS 2.0 looks for the broad consensus across communities.

CONTENTS

01 Introduction

02 Standards, they are great…

03 Multilingual Web LT and ITS 2.0

04 ITS 2.0 interoperability CMS/TMS

05 Online MT Internationalisation

06 Conclusions

Standards, they are great. Everyone should have their own.

- Standards are sometimes produced in excess, making them compete with one another for the same purpose.
- By contrast, new technologies and paradigm shifts that occur in all disciplines require new rules for new needs.
- In this context, the viability of the Web's multilingualism needs a certain level of metadata standards.

Time flies like an arrow

- The multilingual information and knowledge society demands the development, dissemination and adoption of new standards.
- The problem is that the speed of this society does not allow this to take as long as the 'Space Shuttle and the horse's Rear End' did.

Standards help everybody

- They help SMES to:
  - Compete better and faster.
  - Be more compatible, avoiding customer reluctance.
- And help large companies to:
  - Lead the market by leading standards.
  - Facilitate new extensions and features by using standards.
- Open source communities could certainly become open-open, i.e. open source based on open standards.
- And… of course: users.

CONTENTS

01 Introduction

02 Standards, they are great…

03 Multilingual Web LT and ITS 2.0

04 ITS 2.0 interoperability CMS/TMS

05 Online MT Internationalisation

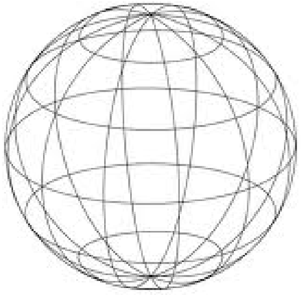
06 Conclusions

MultilingualWeb-LT and ITS 2.0

The **W3C MultilingualWeb-LT Working Group** receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

Linguaserve is a member of MultilingualWeb-LT because:

- Standards help us (as an SME).
- There is no magic button: human language and translation are extremely complex.
- Web content annotation greatly helps to improve results in Multilingual Web Linguistic Technology.



MultilingualWeb metadata requirements

- Information in Web content that is relevant for language technology processing.
- Processes for creating Web content via localization and a content management workflow.
- Language technology applications, tools and resources used in applications that use or support this standard.

ITS 2.0 data categories

- Translate
- Localization Note
- Terminology
- Directionality
- Language Information
- Elements Within Text
- Domain
- Text Analysis
- Locale Filter
- Provenance
- External Resource
- Target Pointer
- Id. Value
- Preserve Space
- Localization Quality Issue
- Localization Quality Rating
- MT Confidence
- Allowed Characters
- Storage Size

Formats supported by ITS 2.0

- ITS 2.0 supports XML-based formats and HTML5, and it is useful for XHTML, and CMS-based ‘deep web’, DITA, DocBook, and mapped to RDF/NIF and XLIFF.

- ITS 2.0 also introduces or modifies important mechanisms like local and global explicit selection rules.

- See http://www.w3.org/TR/its20/

ITS 2.0 implementations

- More than 20 implementations in different areas (see http://www.w3.org/International/its/wiki/Use_cases_-_high_level_summary).

- Two are presented here:
  - Interchange between Content Management System and Translation Management System
  - Content Internationalization and Advanced Machine Translation

- MultilingualWeb-LT has also laid the technical foundations for new business opportunities.

CONTENTS

01 Introduction

02 Standards, they are great…

03 Multilingual Web LT and ITS 2.0

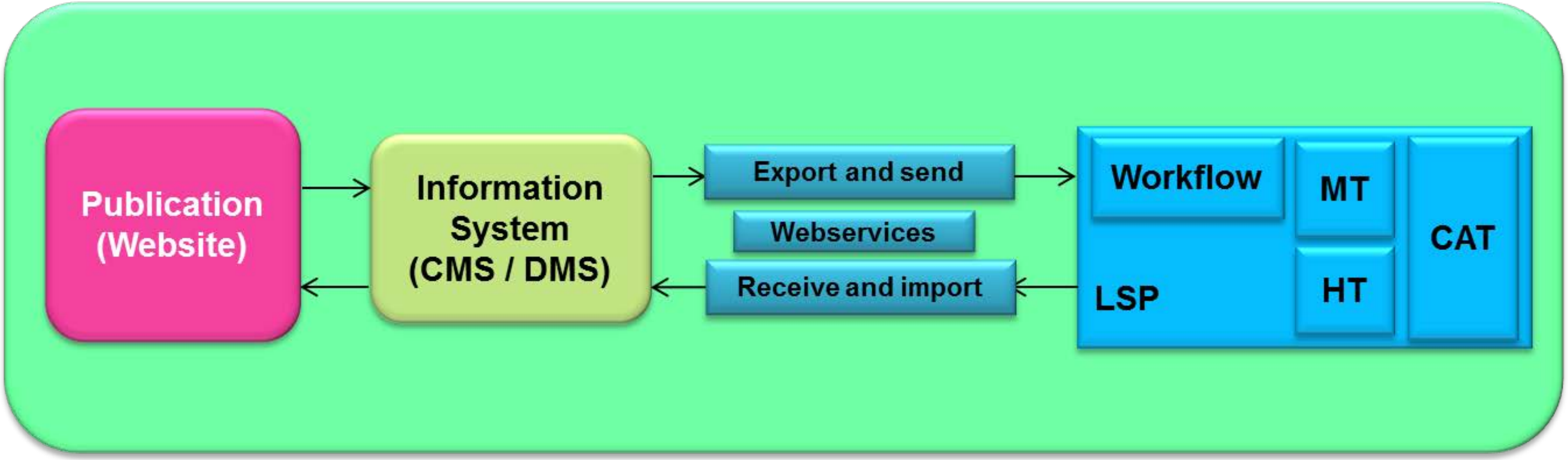
04 ITS 2.0 interoperability CMS/TMS

05 Online MT Internationalisation

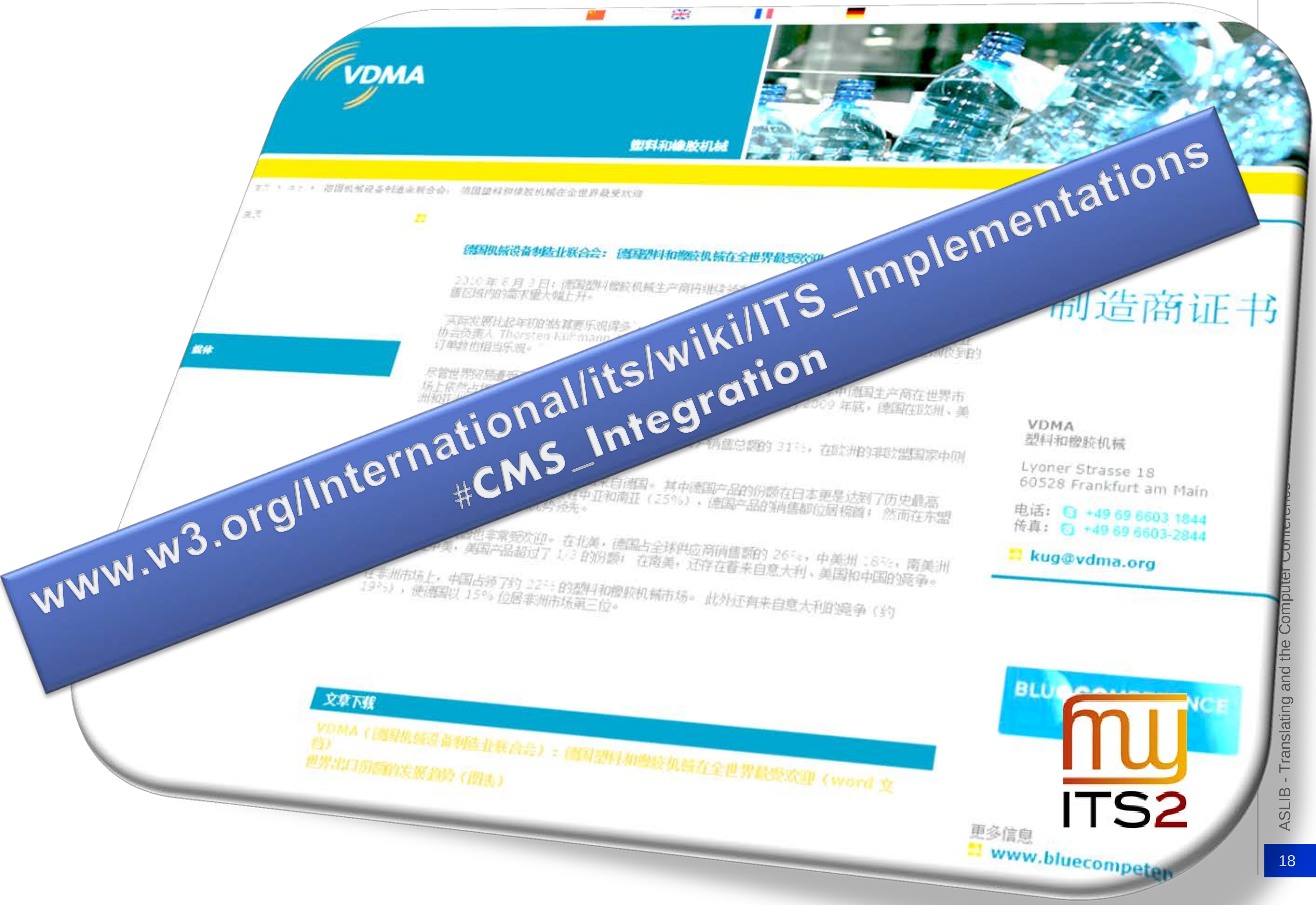
06 Conclusions

linguaserve

Automated off-line translation system (Interoperability)



04. ITS 2.0 interoperability CMS/TMS



德国机械设备制造业联合会：德国塑料和橡胶机械在全世界最受欢迎

2010 年 8 月 3 日：德国塑料橡胶机械生产商将继续……
……区域内的需求量大幅上升。

实际发展比起年初的估算要乐观得多……
协会负责人 Thorsten Kühmann ……
订单数也相当乐观。

尽管世界贸易遭……
场上依然占……
洲和北……

……中德国生产商在世界市
……2009 年底，德国在欧洲、美

……销售总额的 31%，在欧洲的非欧盟国家中则

……来自德国。其中德国产品的份额在日本更是达到了历史最高
……中亚和南亚（25%），德国产品的销售都位居榜首；然而在东盟
……势领先。

……也非常受欢迎。在北美，德国占全球供应商销售额的 26%，中美洲 18%，南美洲
……美，美国产品超过了 1/3 的份额；在南美，还存在着来自意大利、美国和中国的竞争。
在非洲市场上，中国占领了约 22% 的塑料和橡胶机械市场。此外还有来自意大利的竞争（约
19%），使德国以 15% 位居非洲市场第三位。

制造商证书

VDMA
塑料和橡胶机械

Lyoner Strasse 18
60528 Frankfurt am Main

电话： +49 69 6603 1844
传真： +49 69 6603-2844

kug@vdma.org

BLUECOMPETENCE

文章下载

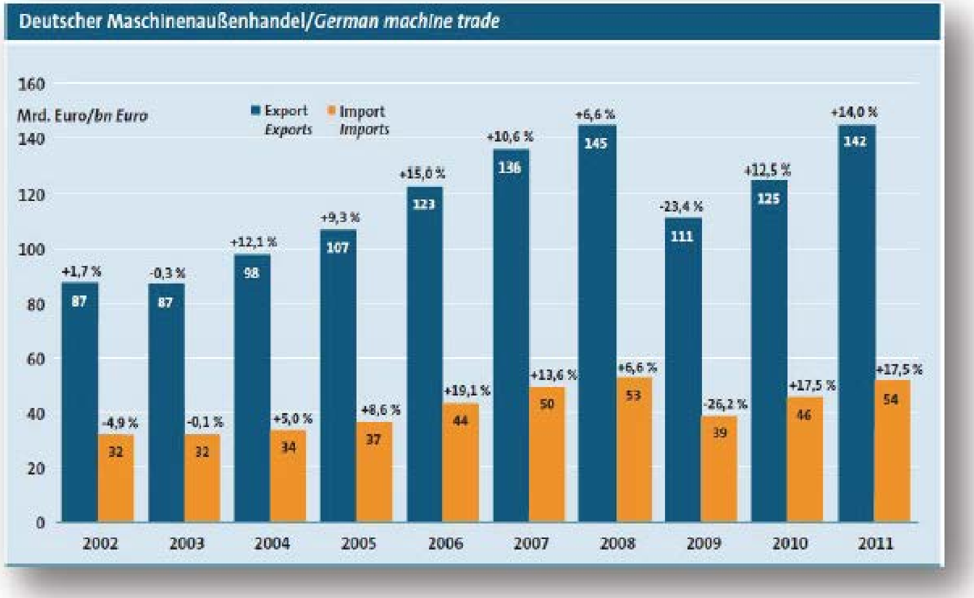
VDMA（德国机械设备制造业联合会）：德国塑料和橡胶机械在全世界最受欢迎（word 文档）
世界出口份额的发展趋势（图表）

更多信息
www.bluecompeten

www.w3.org/International/its/wiki/ITS_Implementations
#CMS_Integration

Use case: VDMA

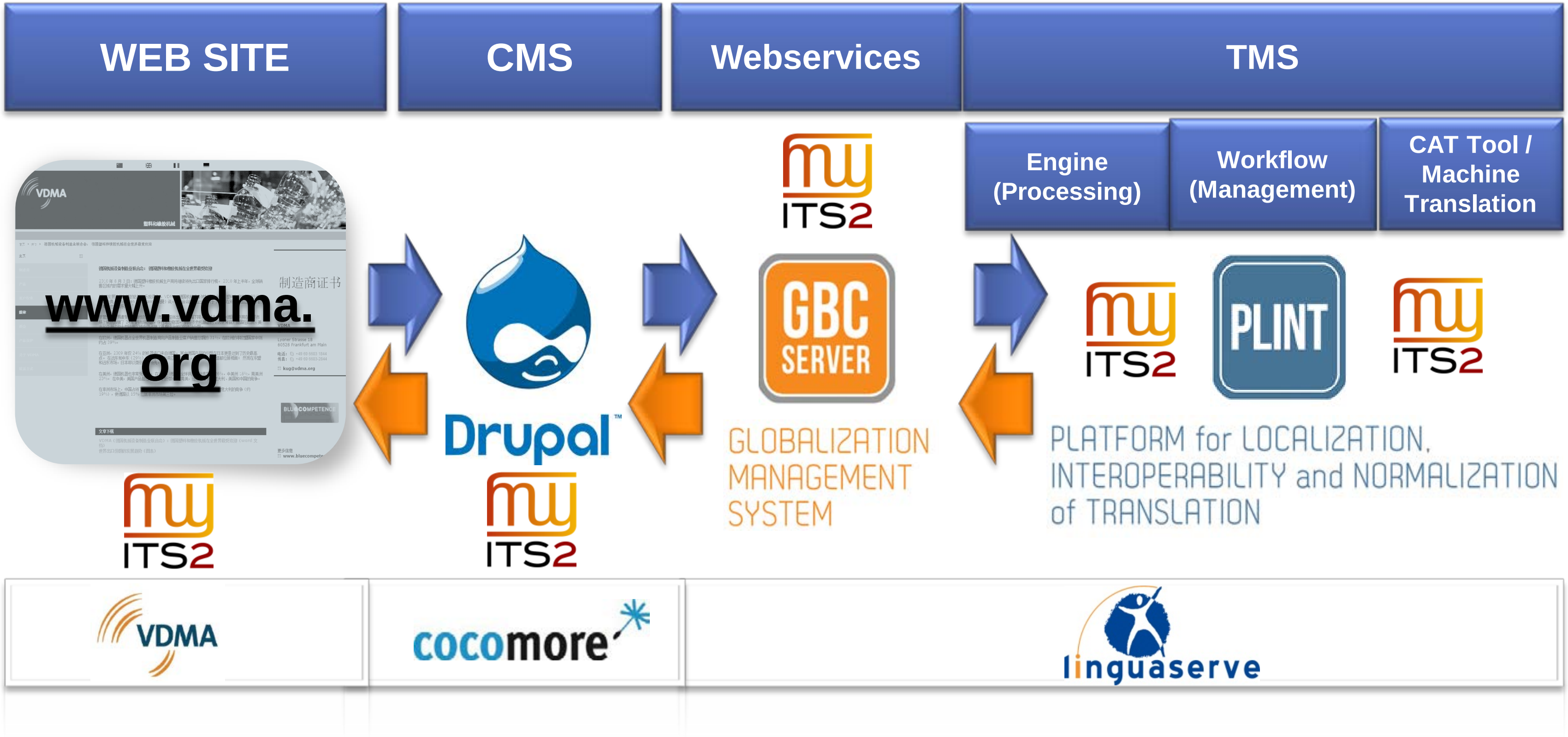
- VDMA: German machinery and plant manufacturers' association
- Largest industrial association in the capital goods industry in Europe (3170 industrial members)
- Highly export-oriented



Use case: Scope

- 150 press releases annotated, processed and translated
- 75,000 words annotated and processed with ITS 2.0
- Using Drupal MLW-LT modules
- ITS 2.0-aware automatic content round-tripping
- Languages: DE > FR, ZH

High level flow



WEB SITE | CMS | Webservices | TMS

Engine (Processing) | Workflow (Management) | CAT Tool / Machine Translation

www.vdma.org

Drupal

GBC SERVER

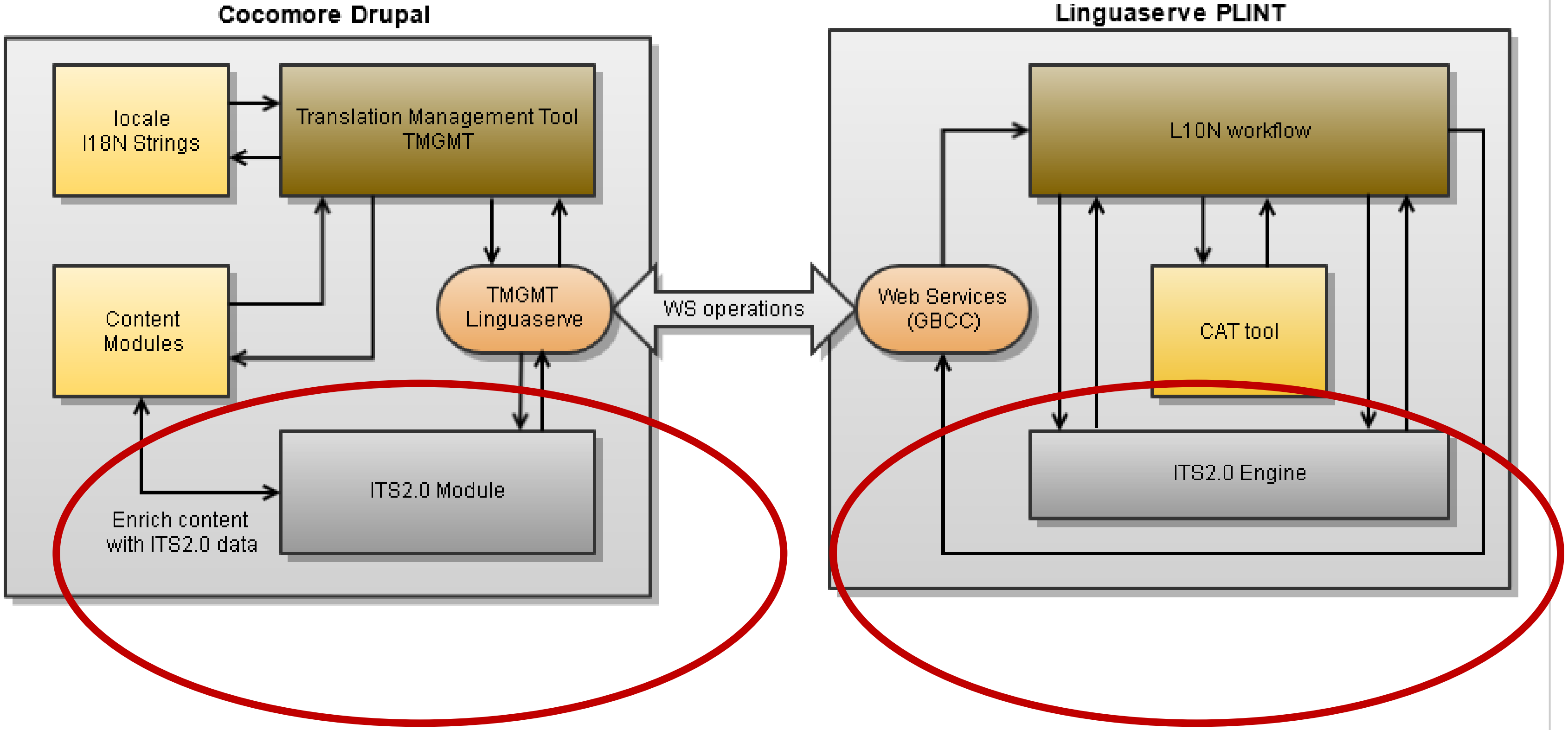
GLOBALIZATION MANAGEMENT SYSTEM

PLINT

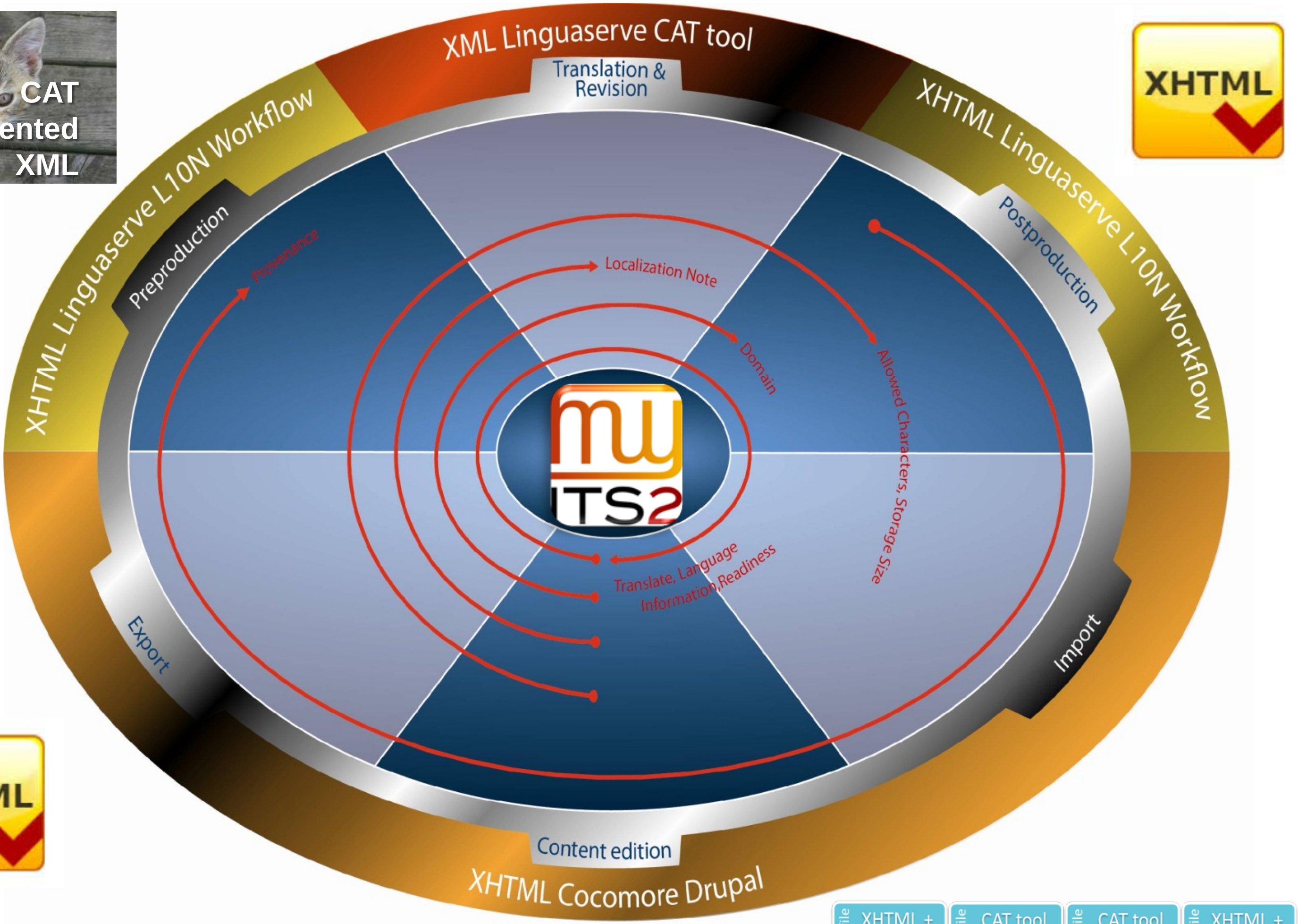
PLATFORM for LOCALIZATION, INTEROPERABILITY and NORMALIZATION of TRANSLATION

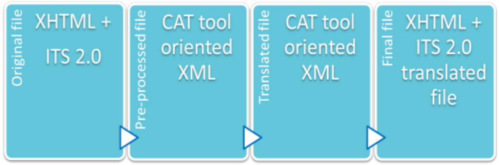
MultilingualWeb-LT

Drupal – GBC Server/PLINT



ITS 2.0 in CMS/TMS processing





XHTML + ITS 2.0

CAT tool oriented XML

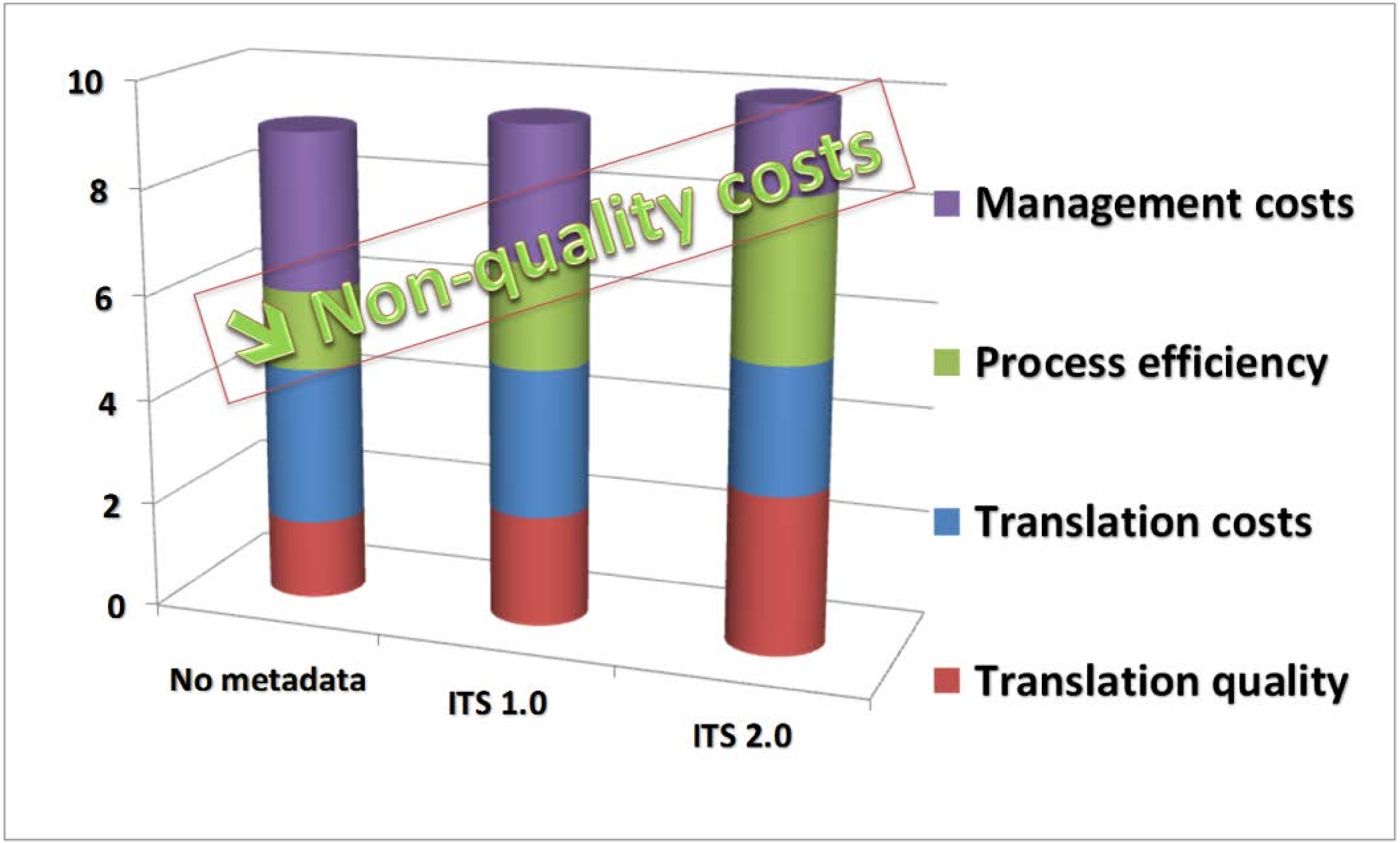
CAT tool oriented XML

XHTML + ITS 2.0 translated file

Test case statistics

- **Data categories**: Translate, Allowed Characters, Localization Note, Storage Size, and Language Information.
- 5,544 tags: 4,700 **manually annotated tags** and the rest **automatically annotated**
- **Density** 39.3 tags per document
- **From** German **into** French and Chinese
- **Other** two data categories were annotated: Provenance and Readiness (ITS 2.0 Extension)
- **Distribution** of data categories: Translate (with value: no) 69.3%; Allowed Characters 11.3%; Provenance 5.4%; Language information 4.3%; Localization Note 3.8%, Storage Size 2.3%; and Readiness 2.3%.

ITS 2.0 impact



Opportunities rise from needs

- Very frequently updated web sites that need efficient multilingual updates and maximum control:
  - Corporate and industry information
  - e-Government
  - e-Commerce
  - Educational web sites
- Highly distributed content creation through the CMS
- Web 2.0 and user content created
  - Applying MT systems for immediacy
- Using ITS 2.0 for multilingual SEO

CONTENTS

01 Introduction

02 Standards, they are great…

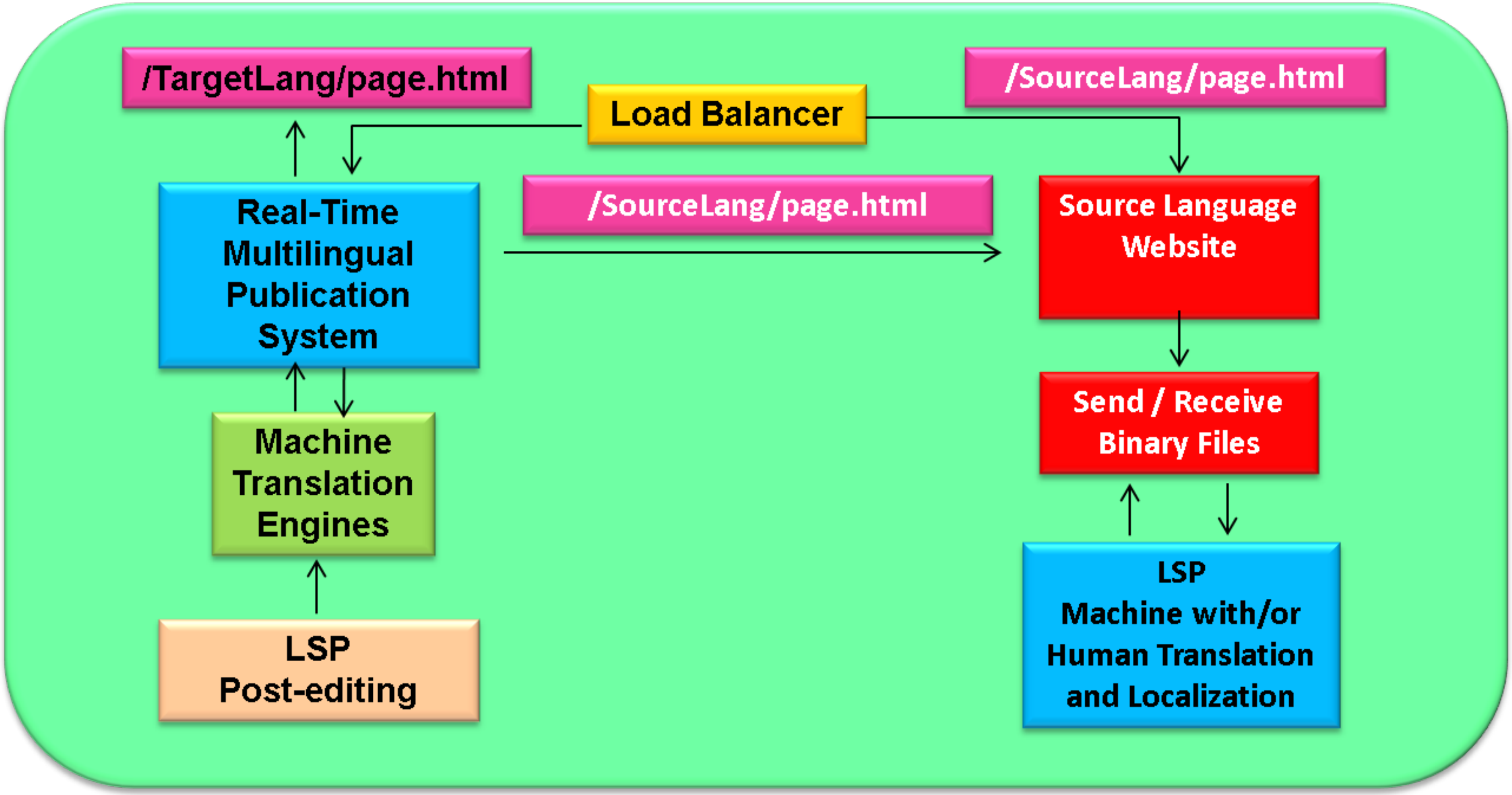
03 Multilingual Web LT and ITS 2.0

04 ITS 2.0 interoperability CMS/TMS

05 Online MT Internationalisation

06 Conclusions

Multilingual Web Publishing System (Real-Time)


/TargetLang/page.html
Load Balancer
/SourceLang/page.html
Real-Time Multilingual Publication System
/SourceLang/page.html
Source Language Website
Send / Receive Binary Files
Machine Translation Engines
LSP Post-editing
LSP Machine with/or Human Translation and Localization

06. Conclusions

www.w3.org/International/its/wiki/ITS_Implementations #Real_Time_Multilingual_Publishing

Use case: the Spanish Tax Agency

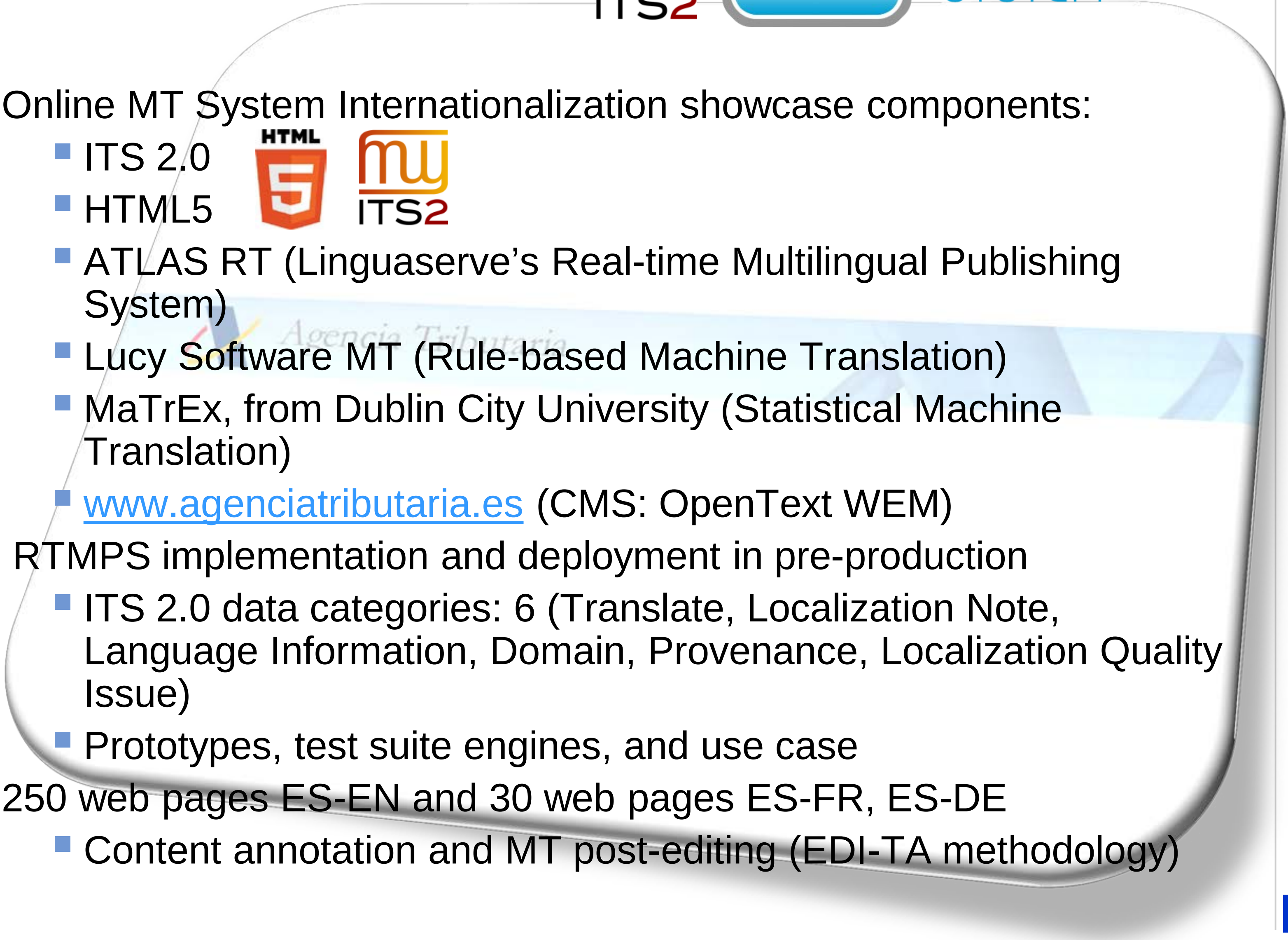
- www.agenciatributaria.es is the user in the “Online MT System” showcase in MLW-LT
- **Spain: General Indicators 2011**
- Spain is a country that is regionally structured into 17 autonomous communities and 2 autonomous cities with **5 co-official languages**
- Population : 47,190,493 inhabitants (**12.2 % foreign residents**)
- **Mission of the Spanish Tax Agency**
- Effective application of Spain’s tax and custom system
- Management of tax resources on behalf of other public administrations when required by Law or Agreements
- **General taxpayer census**

Individual taxpayers:	46,509,231
Companies:	2,674,547
Other organisations:	2,293,939
Total taxpayers:	**51,477,717**

GOBIERNO DE ESPAÑA

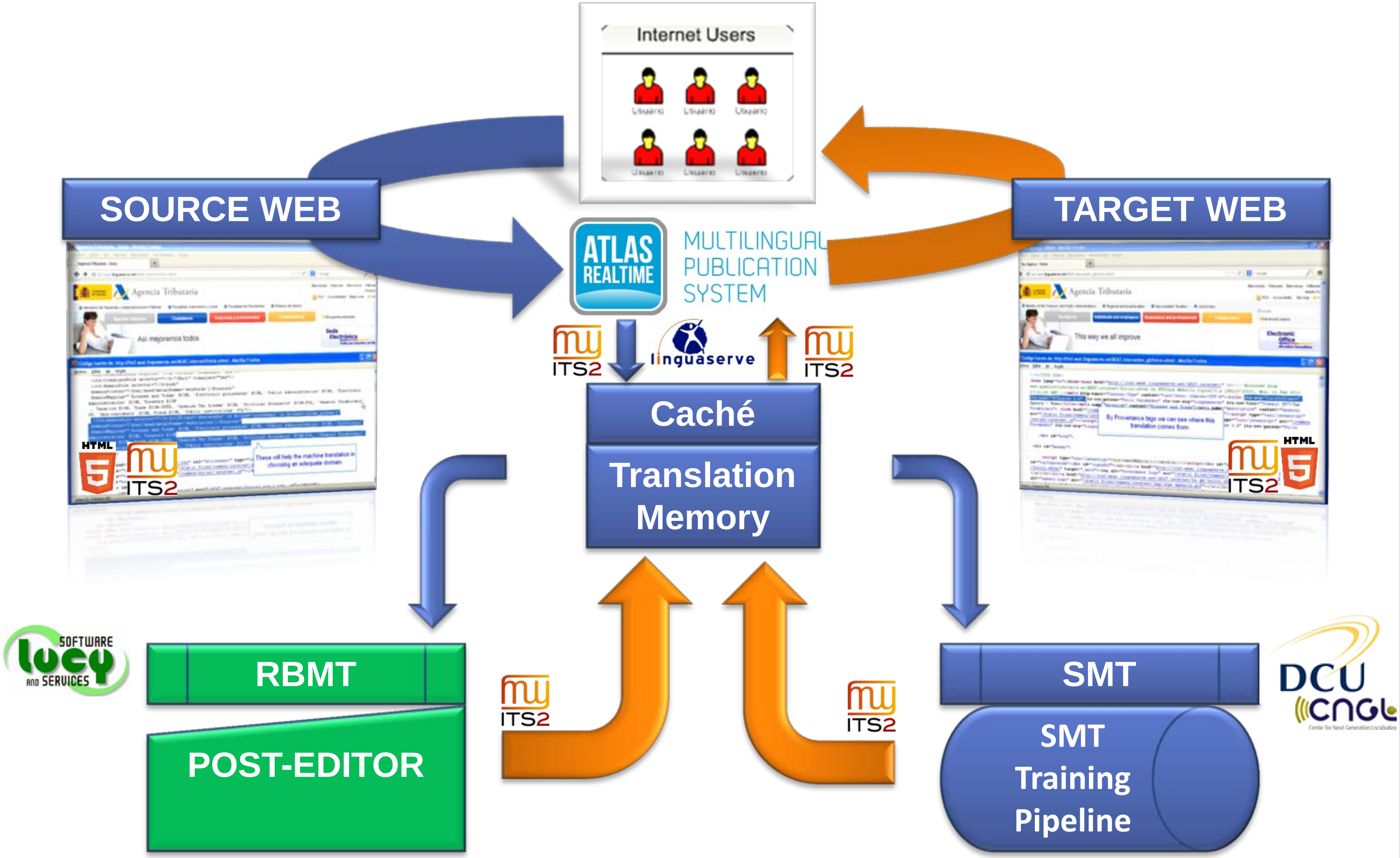
Agencia Tributaria

Use case: Scope

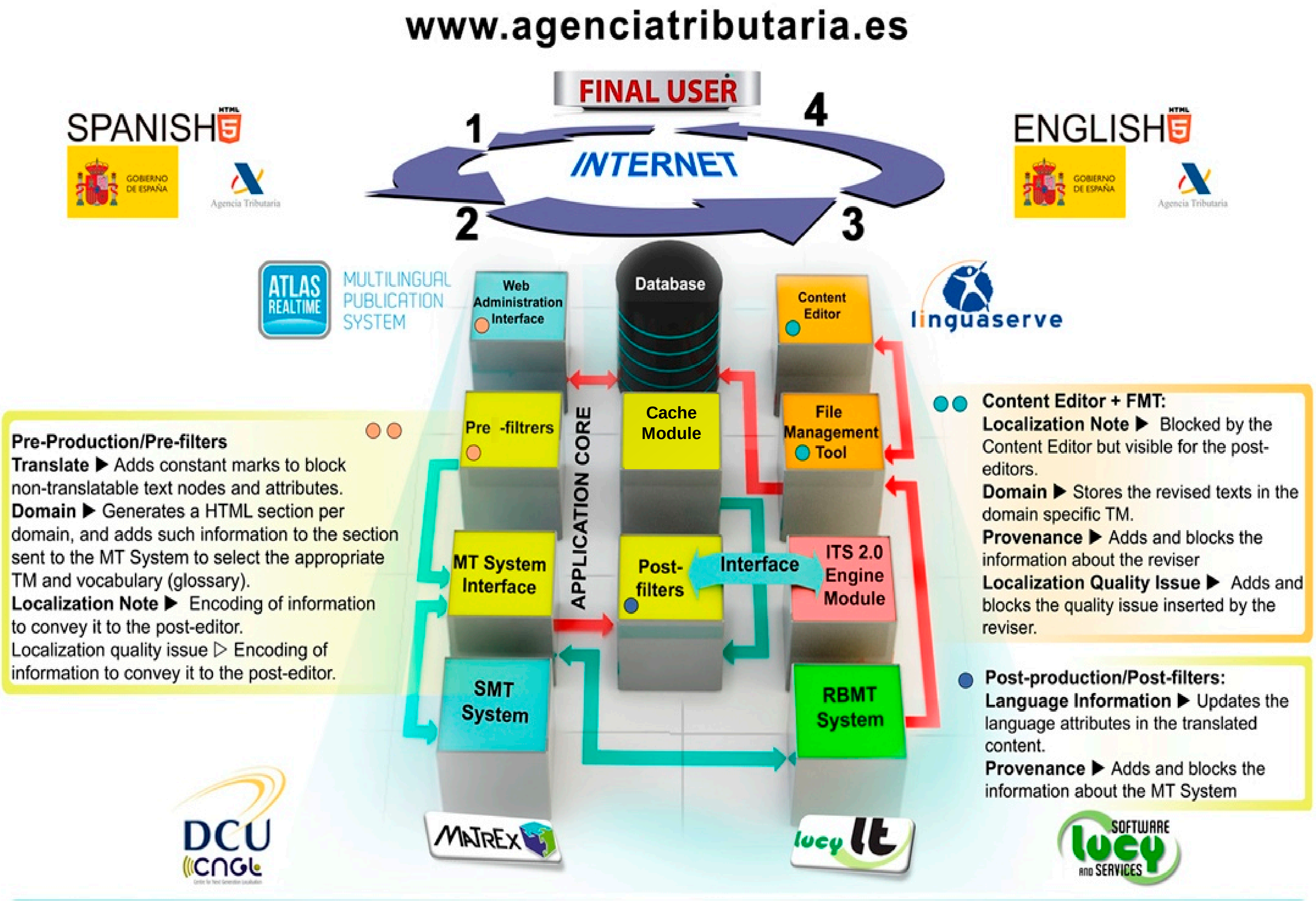


- Online MT System Internationalization showcase components:
  - ITS 2.0
  - HTML5
  - ATLAS RT (Linguaserve's Real-time Multilingual Publishing System)
  - Lucy Software MT (Rule-based Machine Translation)
  - MaTrEx, from Dublin City University (Statistical Machine Translation)
  - www.agenciatributaria.es (CMS: OpenText WEM)
- RTMPS implementation and deployment in pre-production
  - ITS 2.0 data categories: 6 (Translate, Localization Note, Language Information, Domain, Provenance, Localization Quality Issue)
  - Prototypes, test suite engines, and use case
- 250 web pages ES-EN and 30 web pages ES-FR, ES-DE
  - Content annotation and MT post-editing (EDI-TA methodology)

Online MT System I18N



Online MT System I18N

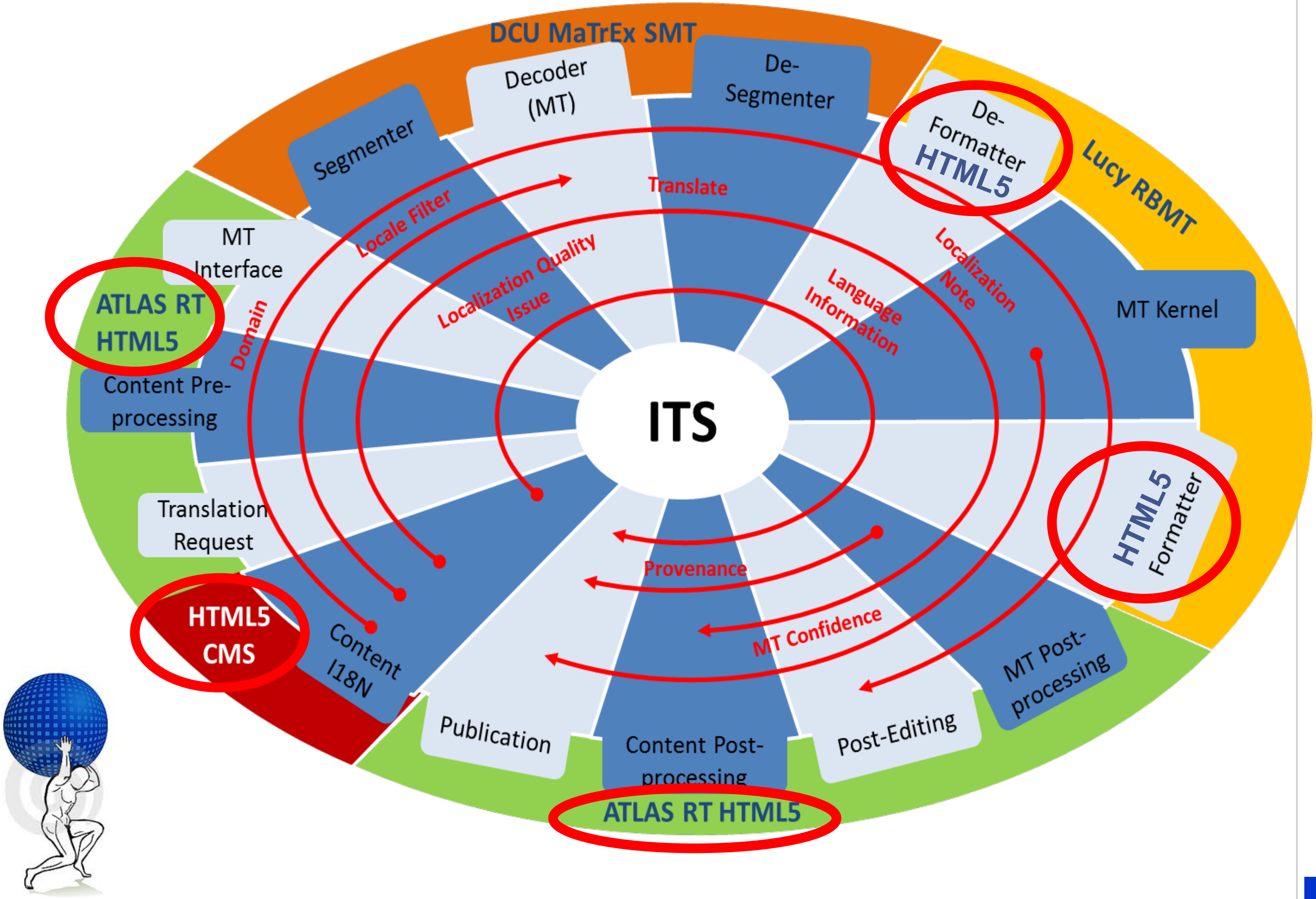


Pre-Production/Pre-filters
Translate ▶ Adds constant marks to block non-translatable text nodes and attributes.
Domain ▶ Generates a HTML section per domain, and adds such information to the section sent to the MT System to select the appropriate TM and vocabulary (glossary).
Localization Note ▶ Encoding of information to convey it to the post-editor.
Localization quality issue ▷ Encoding of information to convey it to the post-editor.

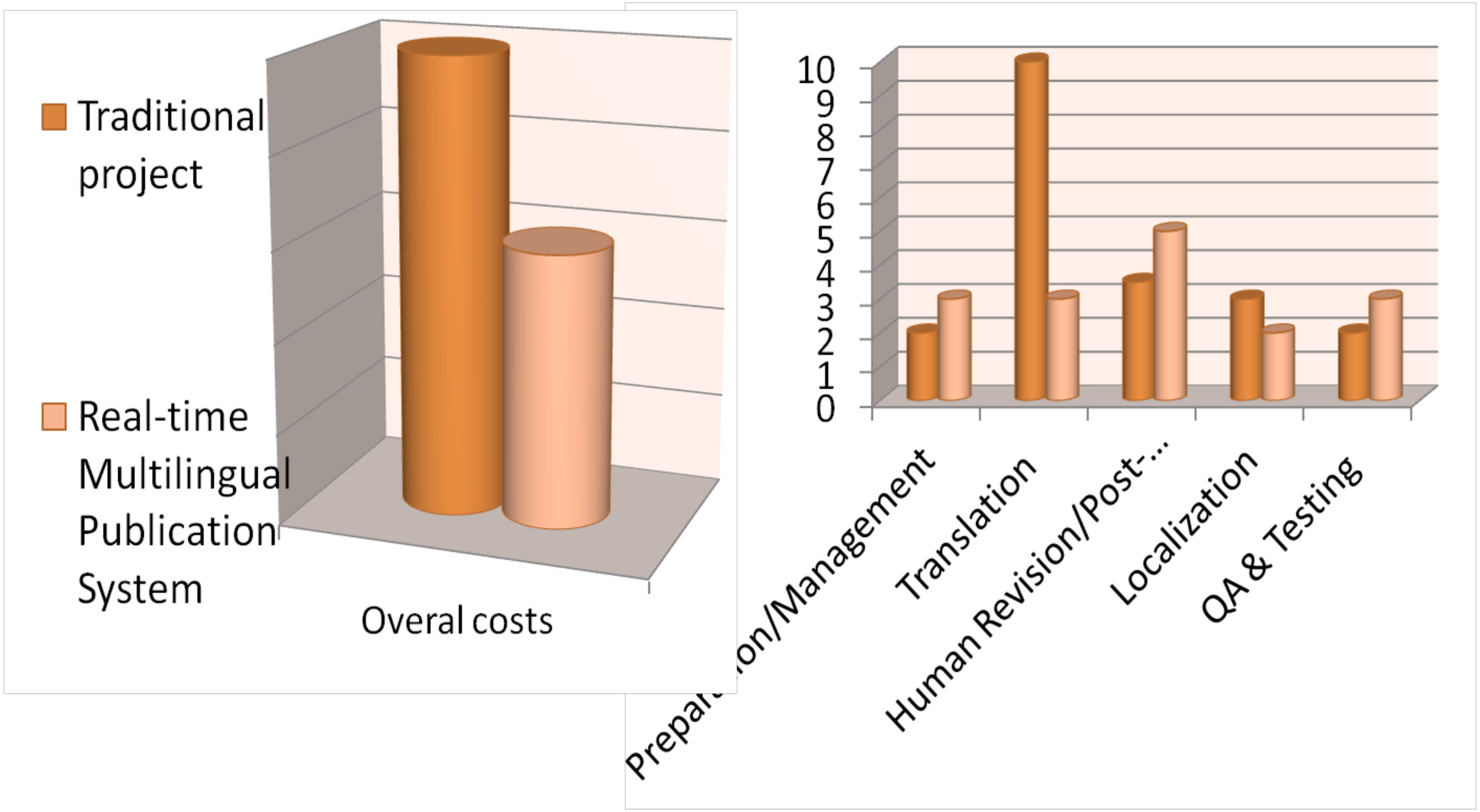
Content Editor + FMT:
Localization Note ▶ Blocked by the Content Editor but visible for the post-editors.
Domain ▶ Stores the revised texts in the domain specific TM.
Provenance ▶ Adds and blocks the information about the reviser
Localization Quality Issue ▶ Adds and blocks the quality issue inserted by the reviser.

Post-production/Post-filters:
Language Information ▶ Updates the language attributes in the translated content.
Provenance ▶ Adds and blocks the information about the MT System

ITS2 in Online MT System I18N



Shifting gears: New cost structure



MLW-LT Online MT Business Case

Strengths

- Lower translation costs (MT + PE) depending on % of post-editing (E.g. 100% post-edited: -30%)
- Management costs: higher setup / lower maintenance (-60% -80%)
- Non-invasive technology
- Real-time or fast post-edition

Weaknesses

- Viability depending on:
- Language combination and MT system output

Recent MT approaches (Hybrids, vertical sectors/users…)

Opportunities

- Web sites with daily high volume updates: E-commerce, Administration, Corporate news and publications, user content generated (social media)
- In house installation for > 1 million words and frequently updated

Threats

- Control, performance and security:

The client might lose control of translation: solved with ITS 2.0

- Real-time performance of MTs
- Security level in shared RTMPS
- Needs pre-editing and post-editing tools (ITS 2.0 and HTML5)

Opportunities rise from needs

- e-Commerce
  - Very high volume and rotation
  - Short texts and repetitive descriptions
    - **Better for MT**
    - **Quicker to post-edit**
  - Very sensitive to ITS 2.0 benefits
- e-Government
  - Controlled language and content policies
- HTML from several CMS and other applications (Content source independent)
- Web 2.0 and user content created
  - GIST translation
  - Immediacy

CONTENTS

01 Introduction

02 Standards, they are great…

03 Multilingual Web LT and ITS 2.0

04 ITS 2.0 interoperability CMS/TMS

05 Online MT Internationalisation

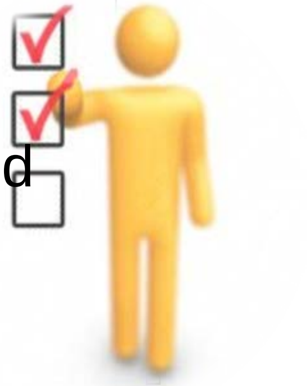
06 Conclusions

linguaserve

ITS 2.0 benefits (I)

- *Translate:* Translatability control from data. E.g. it allows to add “non-translatable” terms to be used by several specific glossaries or MT systems.
- *Localization Note:* Direct communication between webmasters, PMs, translators, and post-editors. When *alert* type, it can be used for triggering certain processes in the Translation Workflow. Activation rules for MT post-editing.
- *Domain:* Automatic selection of CAT/MT terminology, dictionaries, and translation memories.
- *Language Information:* Quality checks to ensure the content’s source language or part of it.
- *Allowed Characters:* Quality check for the target content.

ITS 2.0 benefits (II)



- *Storage Size:* Quality check for both original content and target content. Can also be used for translators' visual control.
- *Provenance:* Identification of agents, possibility to reassign the same translator/reviewer in new versions, and inform the Project Manager. Tracking control in the CMS.
- *Localization Quality Issue:* Quality aspects reported to translation consumer or post-editor.
- *MT Confidence:* Post-editors judge quality of translation.
- *Readiness (ITS 2.0 extension):* Control of processes to be done, date control for availability, delivery and priority.

Win-win business

- More efficient control over the content and faster fine-grain communication between localization chain actors (e.g. webmaster/project manager).
- Localization platforms and format independent.
- Better web and linguistic technology machine/machine interaction.
- Better web and localization human/machine interaction.
- Increasing fully automatic processes and localization expert systems in CMS and TMS.
- Opens up ways for connectors, pre- and post-editing, CAT tools, SEO…
- Time reduction by increasing the efficiency of the process.
- Cost savings in management and translation.

Thank you!

For further information, please contact:

pedro.diez@linguaserve.com